Learning Algorithms for Medical Image Analysis

Matteo Santoro

slipGURU

santoro@disi.unige.it

June 8, 2010

Outline

1. learning-based strategies for quantitative image analysis
2. automatic "annotation" of MR images: the example of synovitis assessment
3. data-driven image representation: *sparse coding* and *dictionary learning*
4. experiments and results

Our Research in Medical Image Analysis

- Definition of new **"imaging biomarkers"** and design of algorithms for their assessment.

  - A biomarker is a "*characteristic that is objectively measured* and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention"

  - An imaging biomarker is, by extension, a biomarker measured from images.

Our Research in Medical Image Analysis

- Definition of new **"imaging biomarkers"** and design of algorithms for their assessment.

  1. Annotation and classification of image parts (i.e. single voxels or 2D/3D patches.)

  2. Non-rigid image registration, with specific interest for novel approaches to deal with discontinuous deformation fields.

Our Research in Medical Image Analysis

- Definition of new **"imaging biomarkers"** and design of algorithms for their assessment.

  1. Annotation and classification of image parts (i.e. single voxels or 2D/3D patches.)

  2. Non-rigid image registration, with specific interest for novel approaches to deal with discontinuous deformation fields.

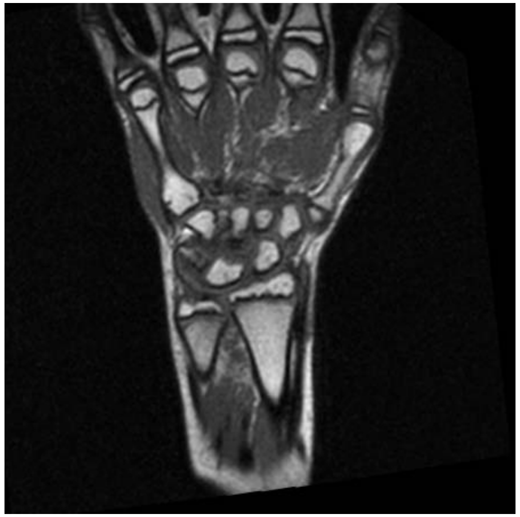
Why learning from examples?

- *It is difficult to "translate" into an explicit algorithm the process of physicians' diagnosis making.*
- *To ask the physicians to annotate images as positive or negative examples is a viable alternative that may lead to implicit learning-based algorithms.*

The Challenges We are Currently Facing

- *quantitative analysis* $\Longrightarrow$ how to design efficient representation schemes to make the analysis more accurate?
- *large collections of 3D images* $\Longrightarrow$ how to design efficient and reliable learning algorithm for large scale problems?
- *weakly annotated data* $\Longrightarrow$ how to combine supervised and unsupervised learning approaches?
- *anatomical constraints* $\Longrightarrow$ how to exploit prior knowledge about known properties of tissue and organs?

Automatic Assessment of Synovial Volume



Setting

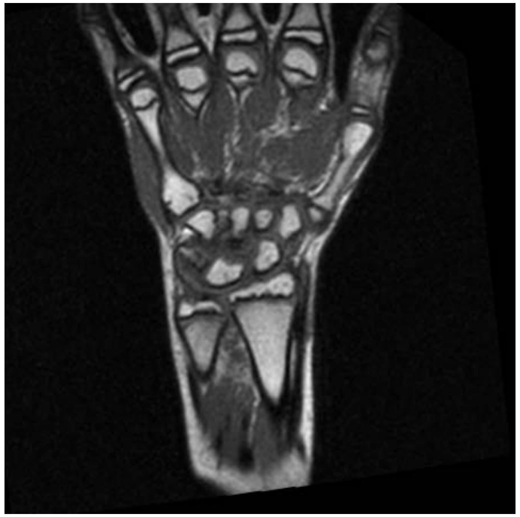
Patients: children under 16, affected by Juvenile Idiopathic Arthritis.

Goal: to measure the volume of the inflamed synovia, and investigate its use as a viable biomarker.

Data: 3D MR images acquired before and after the injection of a contrast medium.

Context: `http://health-e-child.org`

Automatic Assessment of Synovial Volume



The naïve approach

Given n voxels $\mathbf{x}_i \in \mathbb{R}^d$ and the corresponding labels $y_i \in \{+1, -1\}$, the "optimal" solution may be computed as:

$$f(\mathbf{x}) = [K(\mathbf{x}, \mathbf{x}_1), \ldots, K(\mathbf{x}, \mathbf{x}_n)] \cdot \left[(\mathbf{K} + n\tau \mathbf{I})^{-1} \mathbf{y} \right]$$

Automatic Assessment of Synovial Volume

Problems

- nonlinear methods may have storage and performance problems when n becomes large;
- data representation is often heterogeneous (e.g. measurements coming from different modalities):

$$\mathbf{x} \to \phi(\mathbf{x}) = \{\varphi^1, \ldots, \varphi^k\}$$

Multi-cue Voxel Classifier

We look for a more flexible discriminant function:

$$f(\phi) = \sum_{(i,j)\in\mathcal{I}} \alpha_i^j K_i^j(\phi) + b.$$

Multi-cue Voxel Classifier

We look for a more flexible discriminant function:

$$f(\phi) = \sum_{(i,j)\in\mathcal{I}} \alpha_i^j \mathbf{K_i^j(\phi)} + b.$$

Assumptions

The $k \times n$ basis functions

$$K_i^j(\phi) = \exp\left\{-\frac{\|\varphi^j - \varphi_i^j\|^2}{2\sigma_j^2}\right\},$$

measure the similarity between ϕ and one of the exemplar voxels with respect to a specific cue.

Multi-cue Voxel Classifier

We look for a more flexible discriminant function:

$$f(\phi) = \sum_{\mathbf{(i,j)} \in \mathcal{I}} \alpha_i^j K_i^j(\phi) + b.$$

Model Selection

The optimal subset $\mathcal{I}$ of basis functions, on which f depends, may be inferred directly from the data by means of a suitable feature selection algorithm.

Multi-cue Voxel Classifier

We look for a more flexible discriminant function:

$$f(\phi) = \sum_{(i,j)\in\mathcal{I}} \boldsymbol{\alpha}_{\mathbf{i}}^{\mathbf{j}} K_i^j(\phi) + \mathbf{b}.$$

Model Selection

The optimal subset $\mathcal{I}$ of basis functions, on which f depends, may be inferred directly from the data by means of a suitable feature selection algorithm.

Learning Algorithm

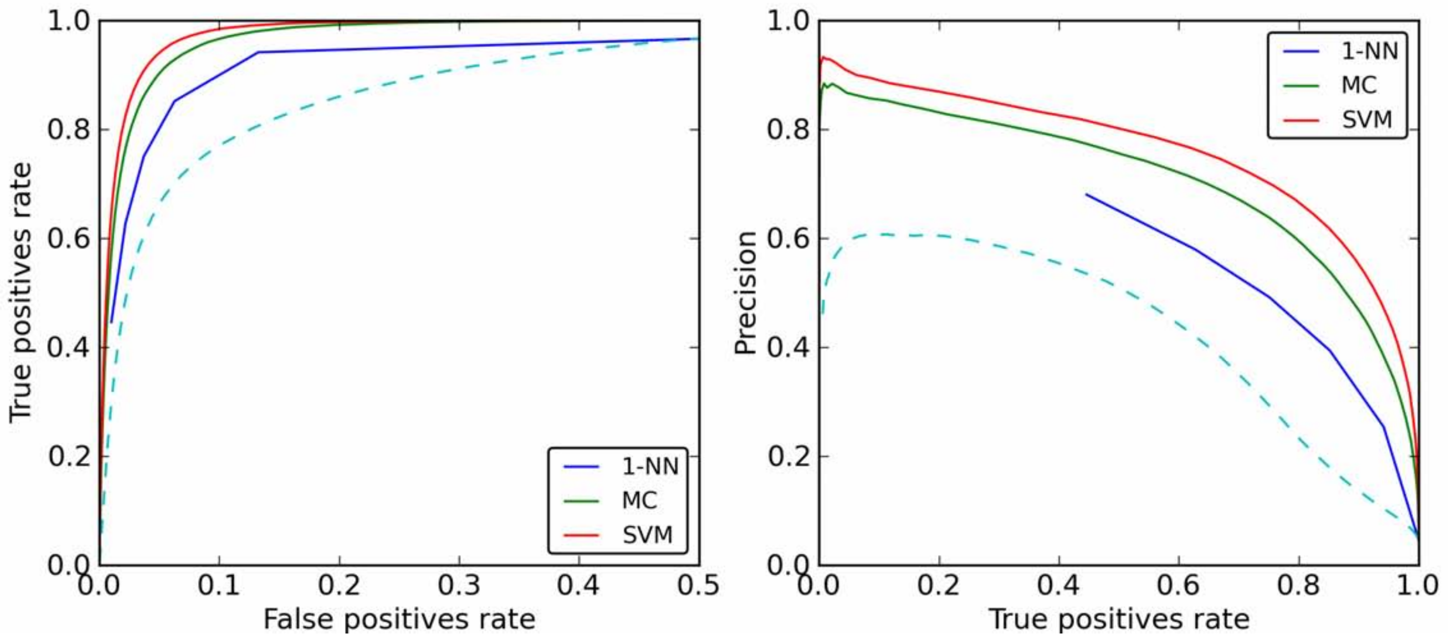
The goal of learning is to to find the optimal affine combination defined by the coefficients α_i^j and b.

Multi-cue Voxel Classifier

We look for a more flexible discriminant function:

$$f(\phi) = \sum_{(i,j)\in\mathcal{I}} \alpha_i^j K_i^j(\phi) + b.$$

Results



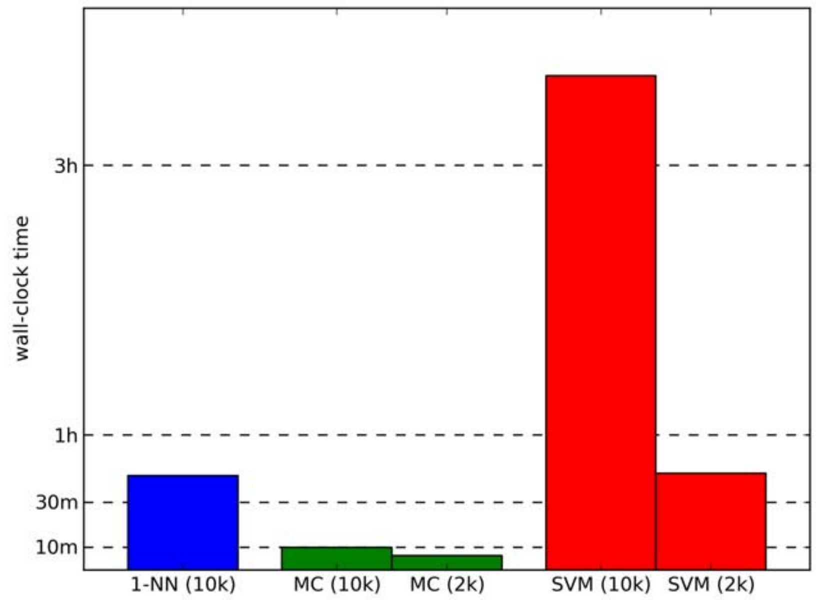
Multi-cue Voxel Classifier

We look for a more flexible discriminant function:

$$f(\phi) = \sum_{(i,j)\in\mathcal{I}} \alpha_i^j K_i^j(\phi) + b.$$

Results

- multi-cue classifier is **15+ times sparser** than SVM,
- and approximately **40 times faster** than SVM.



Multi-cue Voxel Classifier

We look for a more flexible discriminant function:

$$f(\phi) = \sum_{(i,j)\in\mathcal{I}} \alpha_i^j K_i^j(\phi) + b.$$

Results

- excellent accuracy and agreement with the manual measurements
- precision close to the one achieved by manual annotation
- positive preliminary results with both longitudinal and cross-sectional clinical studies.

Data-driven Image Representation

There is empirical evidence that **adaptive** representation schemes may provide effective models.

Data-driven Image Representation

There is empirical evidence that **adaptive** representation schemes may provide effective models.

1. *vector quantisation*:

$$\mathbf{u}_i = j \text{ such that } \|\mathbf{x}_i - \mathbf{D}\mathbf{e}_j\| \leq \|\mathbf{x}_i - \mathbf{D}\mathbf{e}_k\| \forall k \neq j$$

2. *convolution*:

$$\mathbf{u}_i = \mathbf{D}^T \mathbf{x}_i$$

3. *sparse coding*:

$$\min_{\mathbf{u}_i} \left\{ \|\mathbf{x}_i - \mathbf{D}\mathbf{u}_i\|^2 + \lambda \|\mathbf{u}_i\|_1 \right\}$$

Sparse Coding

- A signal may be conveniently represented as the superposition of elementary signals, or *atoms*.
- Over-complete dictionaries (or *frames*) and sparse coding offer more flexibility and are supported by successful applications.
- *Tight frames* ensure that the optimal representation can be recovered by means of inner products of the signal and the dictionary.

Learning the dictionary from data

We investigated the possibility of learning directly from data a dictionary endowed with properties similar to that of tight frames.

Dictionary Learning

Setting

- $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{d \times N}$ is the *input* matrix.
- $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_N] \in \mathbb{R}^{K \times N}$ is the *coding* matrix.

- The goal is to learn:
  1. $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_K] \in \mathbb{R}^{d \times K}$ (the *decoding* or *synthesis* operator), whose columns are the atoms of the dictionary, and
  2. $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_K]^T \in \mathbb{R}^{K \times d}$ (the *encoding* or *analysis* operator), whose rows are filters that can be convolved with an input signal to obtain its encoding.

Dictionary Learning

Objective

We aim at learning both **D** and its dual **C** by minimising:

$$\begin{aligned}\mathcal{E}(\mathbf{D},\mathbf{C},\mathbf{U}) &= \frac{1}{d}\|\mathbf{X}-\mathbf{D}\mathbf{U}\|_F^2 + \frac{\eta}{K}\|\mathbf{U}-\mathbf{C}\mathbf{X}\|_F^2 + \frac{2\tau}{K}\sum_{i=1}^{N}\|\mathbf{u}_i\|_1, \\ s.t. \quad & \|\mathbf{d}_i\|^2, \|\mathbf{c}_i\|^2 \leq 1\end{aligned}$$

where $\tau > 0$ is a regularisation parameter inducing sparsity in **U**, while $\eta \geq 0$ weights the coding error with respect to the reconstruction error.

Dictionary Learning

The Proposed Algorithm: PADDLE

The functional $\mathcal{E}$ is separately convex in each variable, and we adopt a block coordinate descent strategy:

1. *sparse coding* step:
   minimise first with respect to the encoding variables $\mathbf{U}$, and then
2. *dictionary update* step:
   minimise with respect to the dictionary $\mathbf{D}$ and its dual $\mathbf{C}$.

Each step is based on *proximal methods*.

The algorithm has been proved empirically successful, and its convergence towards a critical point of E may be proved.

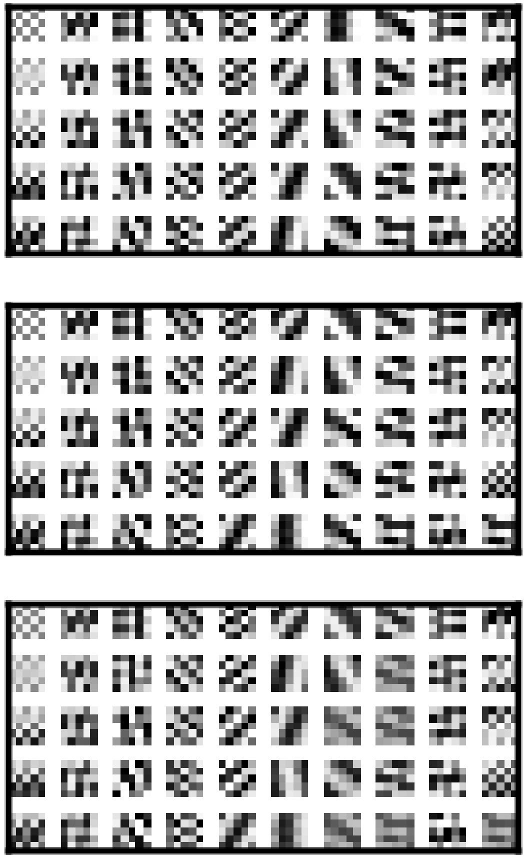
Experimental Assessment

Synthetic Data

Data have been generated as random superpositions of a small number of elements of a tight frame.

The reconstruction error has reached immediately the minimum achievable with the true generating frame.

From top to bottom, we show the original dictionary, the recovered dictionary $\mathbf{D}$ and recovered dual $\mathbf{C}^T$.



Experimental Assessment

Benchmark Datasets

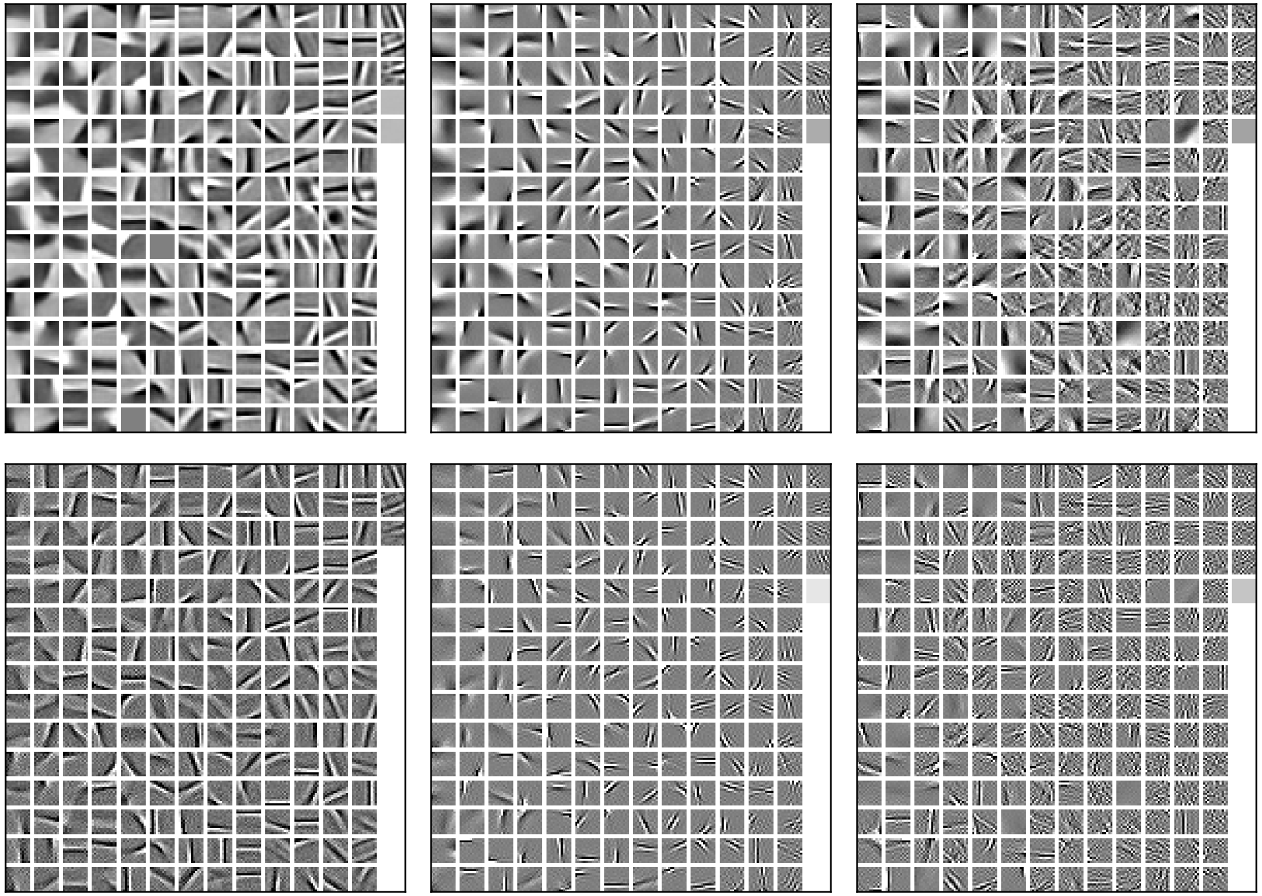
Berkeley segmentation dataset

- we have extracted a random sample of 10^5 patches of size 12×12 from the natural images contained in the dataset
- the reconstruction error achieved at the various level of sparsity have been constantly lower than the reconstruction error achievable with a comparable number of principal components

Experimental Assessment

Benchmark Datasets

Berkeley segmentation dataset



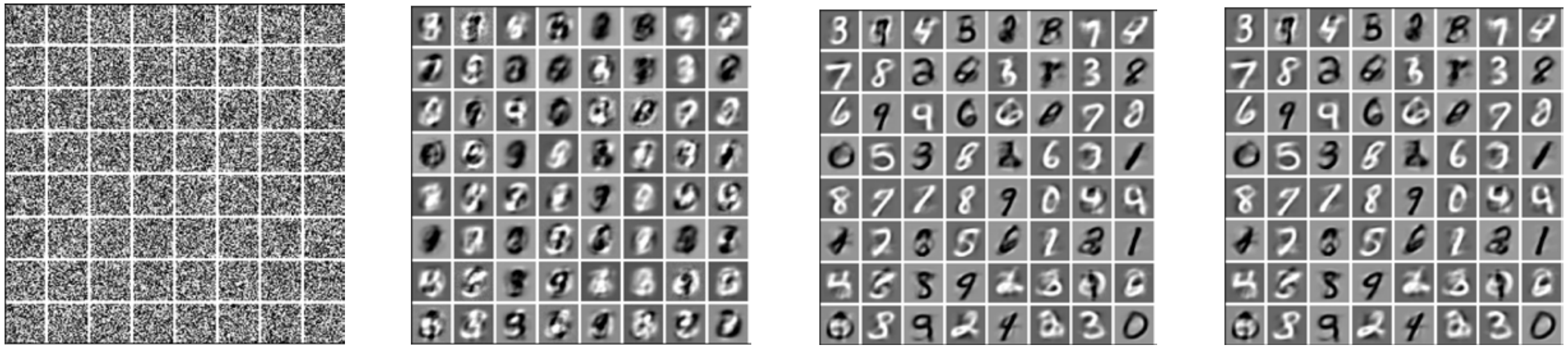
Experimental Assessment

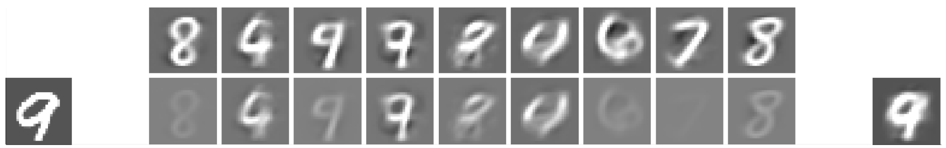
Benchmark Datasets

MNIST dataset

- we have tested the algorithm on the $50,000$ training images consisting of 28×28 quasi binary images of handwritten digits
- we have trained the dictionary with 200 atoms

Initialisation | **After 1 iteration** | **After 4 iterations** | **At convergence**



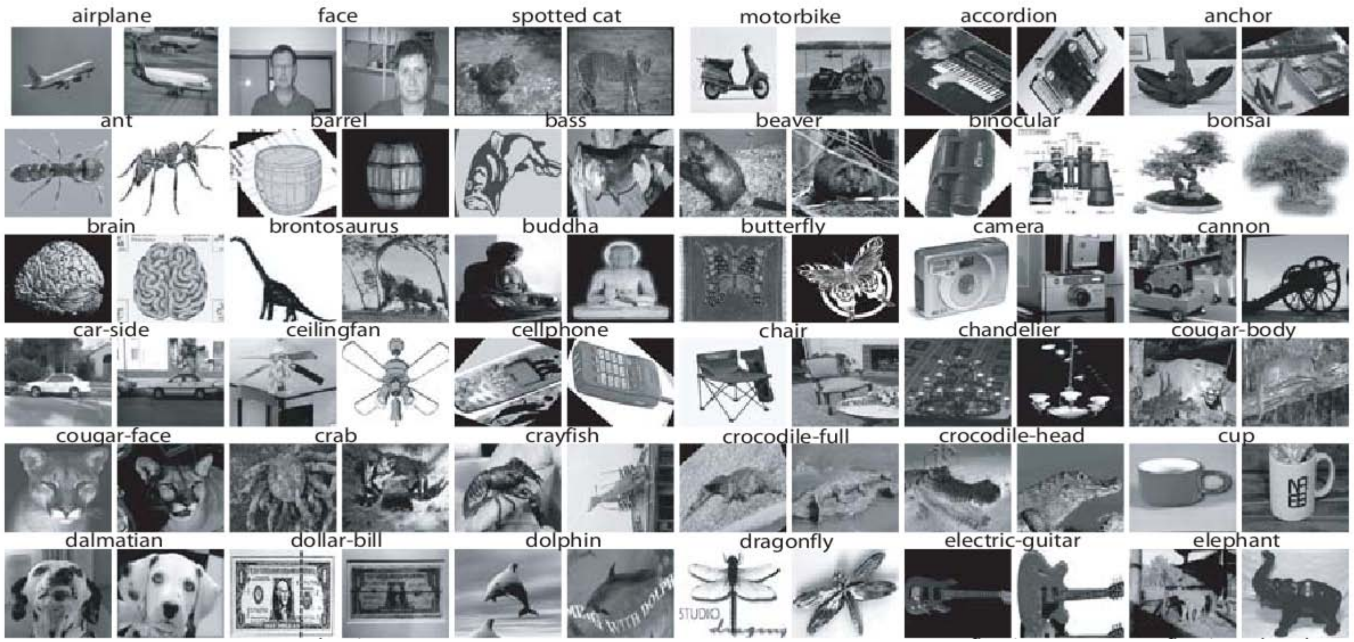


Experimental Assessment

Object Class Identification

Caltech101 Dataset

- we have investigated the discriminative power of the dictionaries **D** and **C** when used to represent the visual content of an image



Experimental Assessment

Object Class Identification

Caltech101 Dataset

- we have investigated the discriminative power of the dictionaries **D** and **C** when used to represent the visual content of an image
- the performance obtained by using **C** is essentially the same as the one obtained with **D**.

	Coding based on D	**Coding based on C**	**State-of-the-art results**
Mean accuracy	0.987	0.984	0.985
Standard deviation	0.008	0.008	0.008

Experimental Assessment

Object Class Identification

Caltech101 Dataset

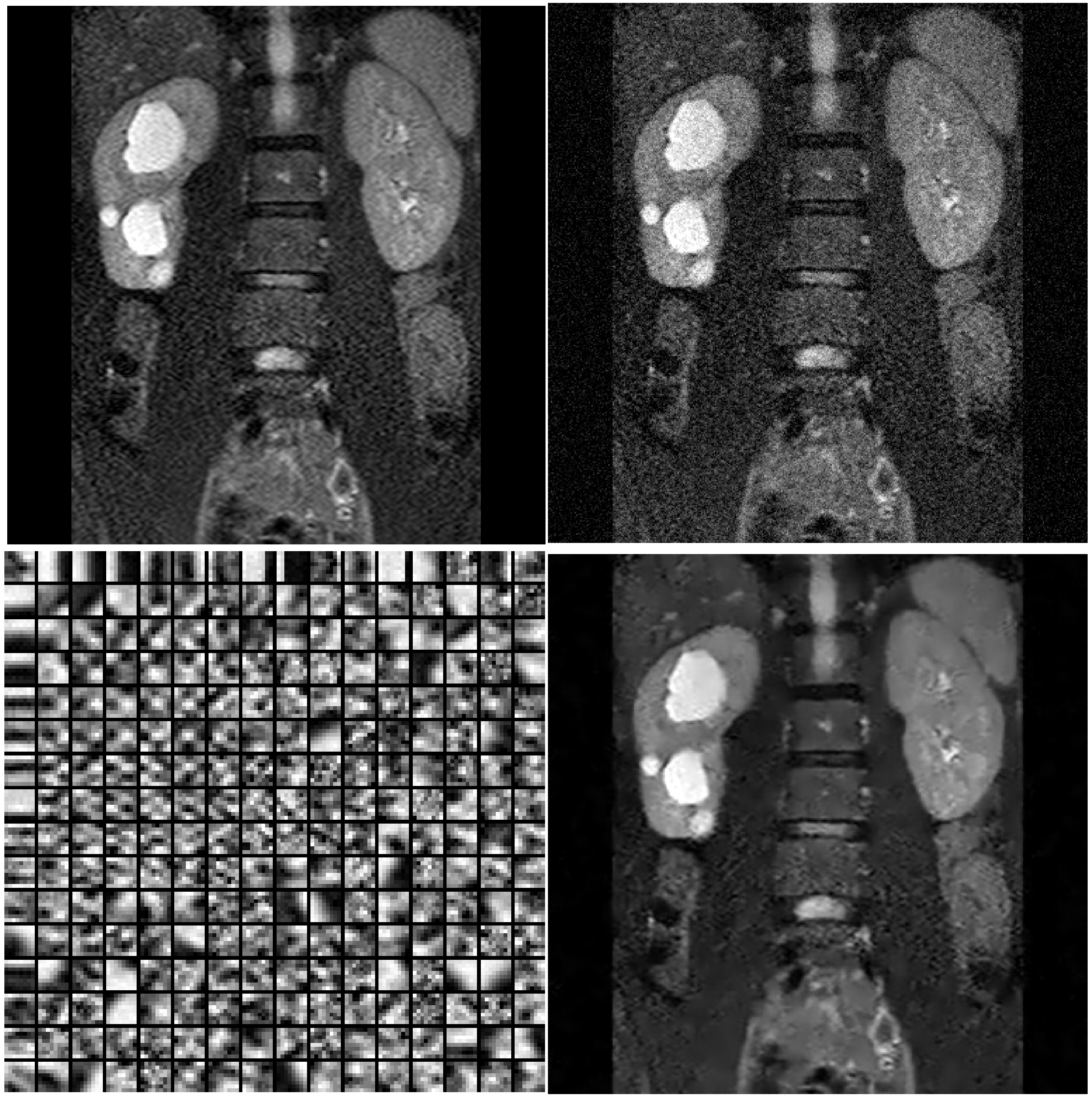
- we have investigated the discriminative power of the dictionaries **D** and **C** when used to represent the visual content of an image
- the performance obtained by using **C** is essentially the same as the one obtained with **D**.

	Coding based on D	**Coding based on C**	**State-of-the-art results**
Mean accuracy	0.987	0.984	0.985
Standard deviation	0.008	0.008	0.008

- the coding of new input images requires only a matrix-vector multiplication.

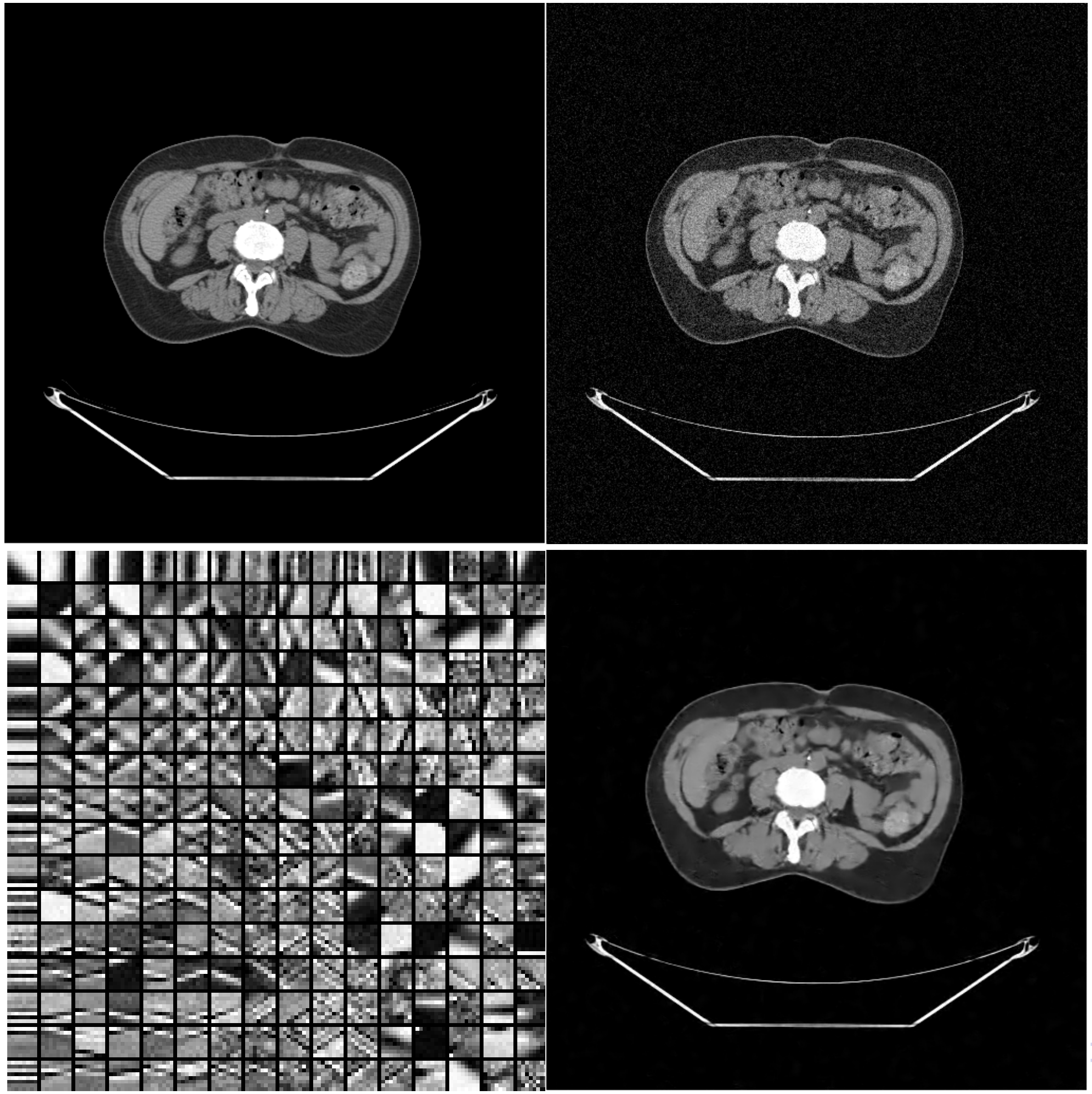
Experimental Assessment

Image Denoising



Experimental Assessment

Image Denoising



References

- C. Basso, M. Santoro, C. Malattia, M.B. Damasio, G. Chiusano, P. Tomà, A. Martini and A. Verri. Quantitative Synovitis Assessment via Automatic 3D MRI Annotation. *Submitted to IEEE Trans Med Imag.*
- C. Basso, M. Santoro, M. B. Damasio, C. Malattia, A. Verri, P. Tomà, and A. Martini. Automatic estimation of inflamed synovial membrane volume in 3D MR images. *Poster at ECR 2010 Scientific Exhibit, 4-8 March 2010, Vienna, Austria.*
- C. Basso, M. Ferrante, M. Santoro, and A. Verri. Automatic Annotation of 3D Multi-Modal MR Images on a Desktop Grid. *MICCAI-Grid Workshop, held at MICCAI 2009, 20-24 September 2009, London, UK.*
- C. Basso, M. Santoro, A. Verri and M. Esposito. Segmentation of Inflamed Synovia in Multi-Modal MRI. *In Proc. of IEEE ISBI 2009, June 28 - July 1 2009.*
- C. Basso, M. Santoro, S. Villa and A. Verri. PADDLE: Proximal Algorithm for Dual Dictionaries LEarning. *Submitted to NIPS 2010.*